Location Patterns of NYC Shooting Incidents

1. Introduction

In this project, I work with the **NYPD Shooting Incident Data (Historic)** dataset. This dataset is publicly available on the Data.gov website and includes every recorded shooting incident in New York City from the beginning of 2006 through the end of 2024.

The purpose of my project is to study the location patterns of shooting incidents in NYC over the past 19 years. By focusing only on the geographic side of the data, I aim to understand where shootings happened more often, and how location patterns have changed over time.

2. Import the Dataset

Below, I import the **NYPD Shooting Incident Data (Historic)** dataset directly from the CSV link I found on Data.gov.

You can use my link https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD directly, or you can go to https://catalog.data.gov/dataset and search for the dataset titled **NYPD Shooting Incident Data (Historic)** for reproducibility purposes.

```
library(tidyverse)
```

```
# Import NYPD Shooting Incident Data
link = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
shootings_raw = read_csv(link)
```

```
## Rows: 29744 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (5): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, Latitude, Longitude
## num   (2): X_COORD_CD, Y_COORD_CD
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Take a brief look at the dataset
glimpse(shootings_raw)
```

```
## Rows: 29,744
## Columns: 21
## $ INCIDENT_KEY            <dbl> 231974218, 177934247, 255028563, 25384540, 726~
## $ OCCUR_DATE              <chr> "08/09/2021", "04/07/2018", "12/02/2022", "11/~
```

```
## $ OCCUR_TIME              <time> 01:06:00, 19:48:00, 22:57:00, 01:50:00, 01:58~
## $ BORO                    <chr> "BRONX", "BROOKLYN", "BRONX", "BROOKLYN", "BRO~
## $ LOC_OF_OCCUR_DESC       <chr> NA, NA, "OUTSIDE", NA, NA, NA, NA, NA, NA, NA,~
## $ PRECINCT                <dbl> 40, 79, 47, 66, 46, 42, 71, 69, 75, 69, 40, 42~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 2, 0, 2, 0, 0, 0, 2, 0, 0, 2, 0~
## $ LOC_CLASSFCTN_DESC      <chr> NA, NA, "STREET", NA, NA, NA, NA, NA, NA, NA, ~
## $ LOCATION_DESC           <chr> NA, NA, "GROCERY/BODEGA", "PVT HOUSE", "MULTI ~
## $ STATISTICAL_MURDER_FLAG <lgl> FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, TRUE, F~
## $ PERP_AGE_GROUP          <chr> NA, "25-44", "(null)", "UNKNOWN", "25-44", "18~
## $ PERP_SEX                <chr> NA, "M", "(null)", "U", "M", "M", NA, NA, "M",~
## $ PERP_RACE               <chr> NA, "WHITE HISPANIC", "(null)", "UNKNOWN", "BL~
## $ VIC_AGE_GROUP           <chr> "18-24", "25-44", "25-44", "18-24", "<18", "18~
## $ VIC_SEX                 <chr> "M", "M", "M", "M", "F", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK", "BLACK", "BLACK", "BLACK", "BLACK", "~
## $ X_COORD_CD              <dbl> 1006343.0, 1000082.9, 1020691.0, 985107.3, 100~
## $ Y_COORD_CD              <dbl> 234270.0, 189064.7, 257125.0, 173349.8, 247502~
## $ Latitude                <dbl> 40.80967, 40.68561, 40.87235, 40.64249, 40.845~
## $ Longitude               <dbl> -73.92019, -73.94291, -73.86823, -73.99691, -7~
## $ Lon_Lat                 <chr> "POINT (-73.92019278899994 40.80967347200004)"~
```

3. Summary and Data Cleaning

After importing the dataset, I first looked at the summary of all columns. Since my project focuses only on location patterns, I only keep the variables that are related to the geography of each incident. These variables are: INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO, PRECINCT, Latitude, and Longitude. All other columns are removed because they are not needed for analyzing location patterns.

```r
# Keeping columns related to where and when each incident occurred
shootings_location = shootings_raw %>%
  select(
    INCIDENT_KEY,
    BORO,
    PRECINCT,
    OCCUR_DATE,
    OCCUR_TIME,
    LATITUDE = Latitude,
    LONGITUDE = Longitude
  )
# Check the dataset
summary(shootings_location)
```

```
##   INCIDENT_KEY           BORO              PRECINCT       OCCUR_DATE
##  Min.   :  9953245   Length:29744       Min.   :  1.00   Length:29744
##  1st Qu.: 67321140   Class :character   1st Qu.: 44.00   Class :character
##  Median :109291972   Mode  :character   Median : 67.00   Mode  :character
##  Mean   :133850951                      Mean   : 65.23
##  3rd Qu.:214741917                      3rd Qu.: 81.00
##  Max.   :299462478                      Max.   :123.00
##
##    OCCUR_TIME                LATITUDE       LONGITUDE
##  Min.   :00:00:00.000000   Min.   :40.51   Min.   :-74.25
##  1st Qu.:03:30:45.000000   1st Qu.:40.67   1st Qu.:-73.94
##  Median :15:15:00.000000   Median :40.70   Median :-73.91
```

```
##  Mean   :12:46:10.874798   Mean   :40.74   Mean   :-73.91
##  3rd Qu.:20:44:00.000000   3rd Qu.:40.83   3rd Qu.:-73.88
##  Max.   :23:59:00.000000   Max.   :40.91   Max.   :-73.70
##                            NA's   :97      NA's   :97
```

After checking the summary of the shootings_location dataset, I found that several changes were needed to clean the data and prepare it for analysis. I convert BORO and PRECINCT into factors because they both are categorical variables. I also convert OCCUR_DATE to a proper date format so I can look at long-term patterns from 2006 to 2024. Additionally, I remove 97 missing values of LATITUDE and LONGITUDE because these incidents cannot be placed on the map or used in any location analysis. After applying these cleaning steps, the dataset is ready for analyzing location patterns in New York City.

```
shootings_clean = shootings_location %>%
  mutate(
    # Change BORO to factor
    BORO = as.factor(BORO),
    # Change PRECINCT to factor
    PRECINCT = as.factor(PRECINCT),
    # Proper date types
    OCCUR_DATE = mdy(OCCUR_DATE)
  ) %>%
  # Remove missing data
  filter(
    !is.na(LATITUDE),
    !is.na(LONGITUDE)
  )

# Check summary of cleaned dataset
summary(shootings_clean)
```

```
##   INCIDENT_KEY                  BORO          PRECINCT       OCCUR_DATE
##  Min.   :  9953245   BRONX        : 8810   75     : 1676   Min.   :2006-01-01
##  1st Qu.: 67109564   BROOKLYN     :11655   73     : 1557   1st Qu.:2009-10-24
##  Median : 95318519   MANHATTAN    : 3953   67     : 1287   Median :2014-03-11
##  Mean   :133385135   QUEENS       : 4414   44     : 1159   Mean   :2014-10-20
##  3rd Qu.:214513240   STATEN ISLAND:  815   79     : 1070   3rd Qu.:2020-06-23
##  Max.   :299462478                        47     : 1047   Max.   :2024-12-31
##                                           (Other):21851
##    OCCUR_TIME                LATITUDE       LONGITUDE
##  Min.   :00:00:00.00000   Min.   :40.51   Min.   :-74.25
##  1st Qu.:03:30:00.00000   1st Qu.:40.67   1st Qu.:-73.94
##  Median :15:15:00.00000   Median :40.70   Median :-73.91
##  Mean   :12:45:40.83651   Mean   :40.74   Mean   :-73.91
##  3rd Qu.:20:43:00.00000   3rd Qu.:40.83   3rd Qu.:-73.88
##  Max.   :23:59:00.00000   Max.   :40.91   Max.   :-73.70
##
```

4. Borough-Level Analysis

In this section, I look at shooting incidents at the borough level to understand the overall location patterns across New York City.

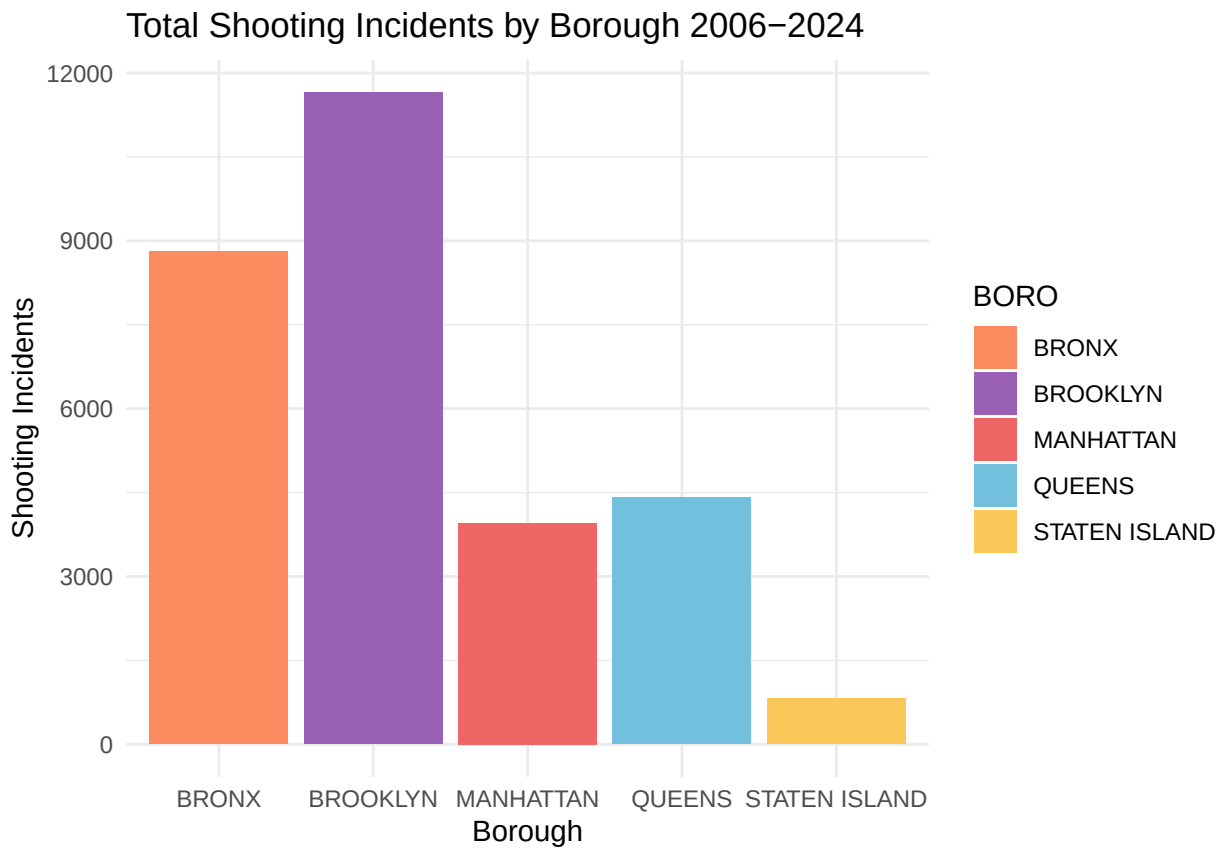
4.1 Total Shooting Incidents by Borough

First, I calculate the total number of shooting incidents for each borough from 2006 to 2024 and make a bar chart. This helps me compare how shooting activity is distributed across the five boroughs.

```
# Count total incidents by borough
borough_total = shootings_clean %>%
  group_by(BORO) %>%
  summarise(total_shootings = n())

# Bar Chart
my_palette = c(
  "BRONX"         = "#FC8B60",
  "BROOKLYN"      = "#9A60B4",
  "MANHATTAN"     = "#EE6666",
  "QUEENS"        = "#73C0DE",
  "STATEN ISLAND" = "#FAC859"
)

ggplot(borough_total, aes(x = BORO, y = total_shootings, fill = BORO)) +
  geom_col() +
  scale_fill_manual(values = my_palette) +
  labs(
    title = "Total Shooting Incidents by Borough 2006-2024",
    x = "Borough",
    y = "Shooting Incidents"
  ) +
  theme_minimal()
```



This bar chart shows the total number of shooting incidents in each borough from 2006 to 2024. Brooklyn has the highest number of shootings during this period, followed by the Bronx. Queens and Manhattan have fewer incidents, and Staten Island has the lowest count. This gives me a general idea of where experienced more shooting activities in New York City.

4.2 Trends in NYPD Shooting Incidents by Borough 2006-2024

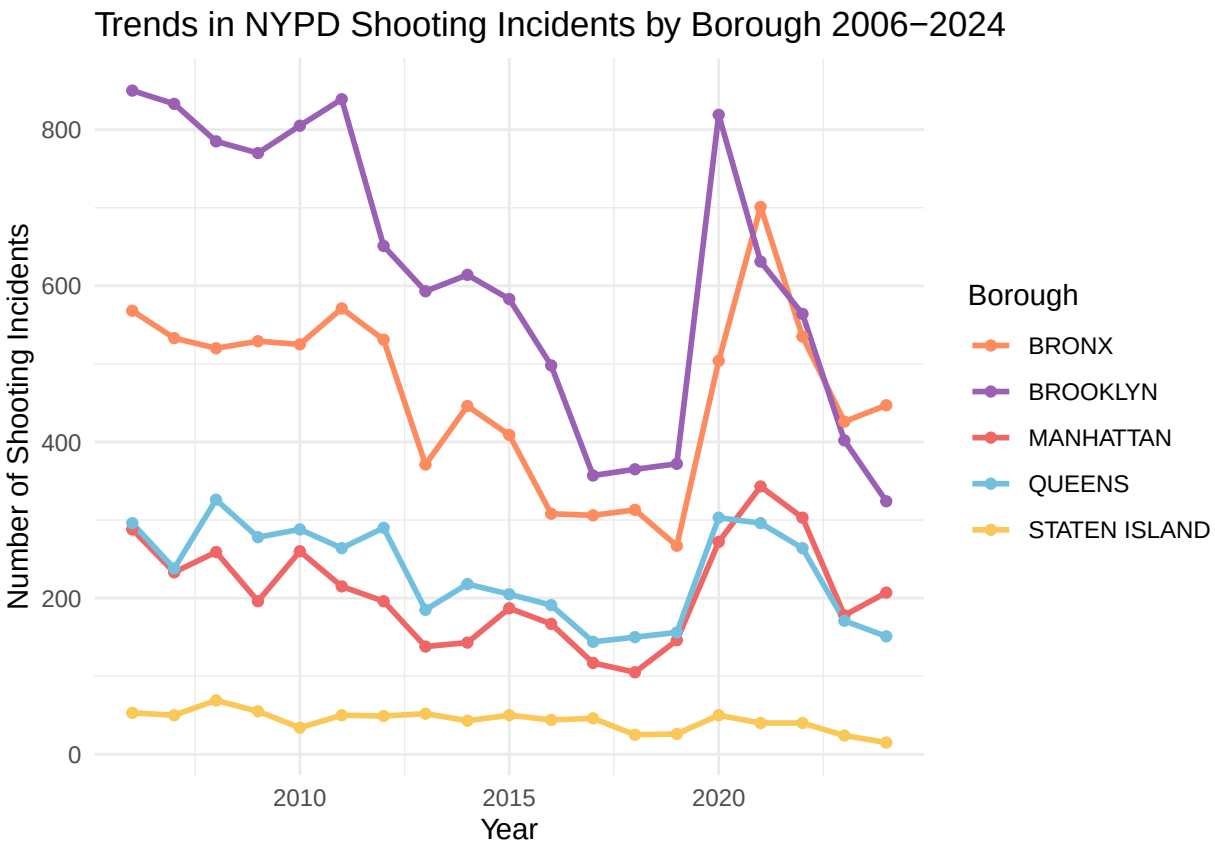
Next I want to take a look at shooting trends over time for each of the five boroughs. To do this, I first extracted the year from each incident date and then calculated the total number of shootings per borough per year. I plotted these counts on a line chart, which clearly shows how shooting activity has changed over the past 19 years in each borough. This visualization makes it easy to compare long-term trends and identify boroughs with consistently high or rising shooting incidents.

```
# Add a year column
shootings_year = shootings_clean %>%
  mutate(YEAR = year(OCCUR_DATE))

# Count shootings per borough per year
shootings_by_boro_year = shootings_year %>%
  group_by(BORO, YEAR) %>%
  summarize(count = n(), .groups = "drop")

# Plot the trend
ggplot(shootings_by_boro_year, aes(x = YEAR, y = count, color = BORO)) +
  geom_line(size = 1) +
  geom_point(size = 1.5) +
```

```
  scale_color_manual(values = my_palette) +
  labs(
    title = "Trends in NYPD Shooting Incidents by Borough 2006-2024",
    x = "Year",
    y = "Number of Shooting Incidents",
    color = "Borough"
  ) +
  theme_minimal()
```



The Chart above shows how shooting incidents have changed in each of the five boroughs from 2006 to 2024. Overall, Brooklyn and Bronx consistently have the highest number of shootings. Queens and Manhattan have moderate levels of shootings compared to Brooklyn and Bronx. Those four boroughs followed similar patterns over the years: a downward trend from around 2012 to 2018, a noticeable increase from 2019 to 2020, and then a continuing drop in numbers after that. Staten Island has the lowest number of incidents every year, with very little change over time.

5. Precinct-Level Analysis

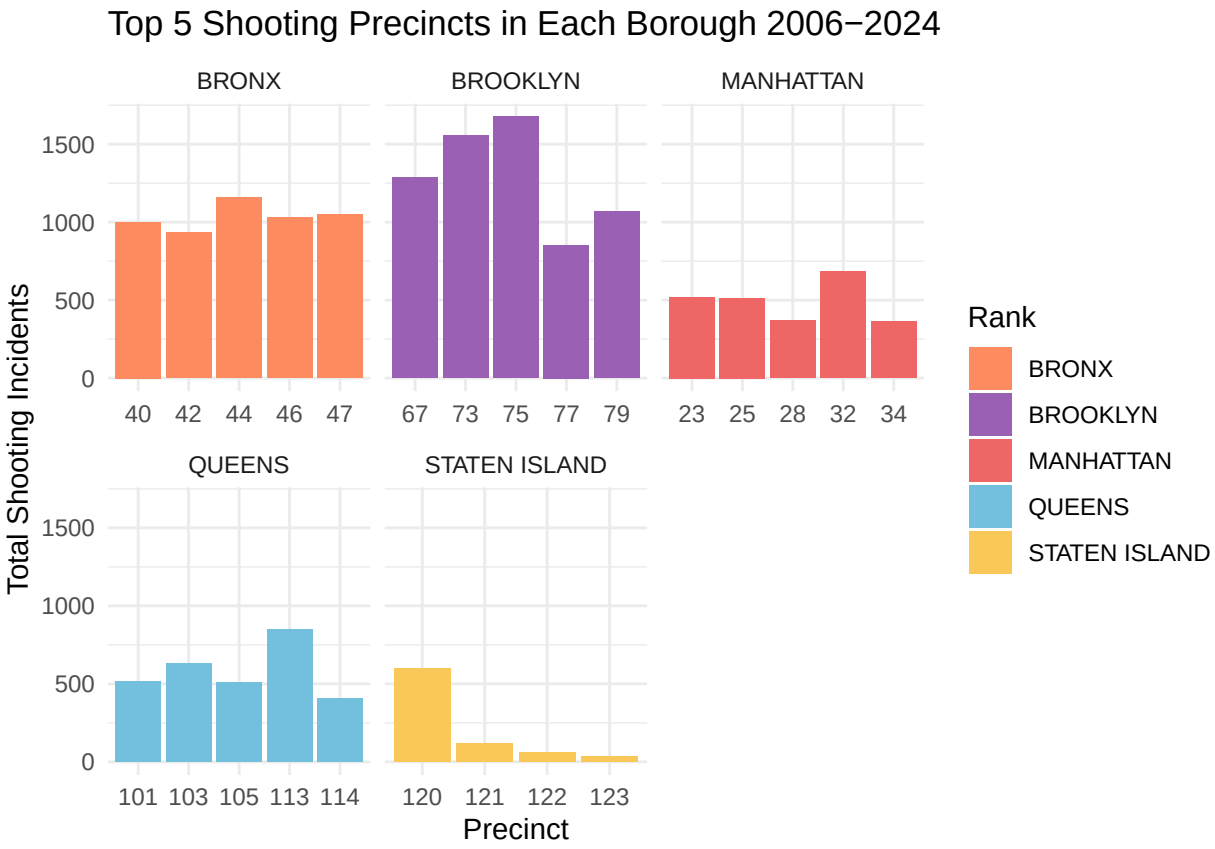
In the previous section, I looked at shooting patterns at the borough level. However, each borough is made up of many precincts, and shooting incidents are not evenly spread inside a borough. To see the location patterns in more detail, I now zoom in to the precinct level.

5.1 Top 5 Shooting Precincts in Each Borough 2006-2024

To focus on the areas that matter most, I look at the top 5 precincts with the highest number of shootings in each borough. This gives a clearer picture of where shootings have been most concentrated from 2006 to 2024, without making the analysis too complicated or overwhelming.

```
# create a table shows top 5 precincts per borough
top5_precinct = shootings_clean %>%
  group_by(BORO, PRECINCT) %>%
  summarize(Total_Shootings = n(), .groups = "drop") %>%
  group_by(BORO) %>%
  slice_max(order_by = Total_Shootings, n = 5, with_ties = FALSE) %>%
  ungroup()

# Draw a bar chart to make it more visualized
ggplot(top5_precinct, aes(x = PRECINCT, y = Total_Shootings, fill= BORO)) +
  geom_col() +
  facet_wrap(~BORO, scales = "free_x") +
  scale_fill_manual(values = my_palette) +
  labs(
    title = "Top 5 Shooting Precincts in Each Borough 2006-2024",
    x = "Precinct",
    y = "Total Shooting Incidents",
    fill = "Rank"
  ) +
  theme_minimal()
```



The chart shows the five precincts with the highest number of shooting incidents in each borough from 2006 to 2024. Brooklyn and the Bronx each have several precincts with very high totals, which matches what we saw at the borough level. In Brooklyn, Precincts 67, 73, and 75 stand out as the main hotspots. In the Bronx, precincts 44, 46, and 47 have the highest shooting activity. Manhattan and Queens also have their own top precincts, such as Manhattan's 32nd Precinct and Queens' 113th Precinct, but their totals are lower compared to the major hotspots in Brooklyn and the Bronx. Staten Island has the lowest numbers overall, and only Precinct 120 shows a noticeable count. This plot shows that shooting incidents are not spread evenly across precincts. Instead, each borough has a few precincts with much higher totals than the others.

6. Model

After exploring the borough-level patterns and the top precincts within each borough, I want to confirm whether the differences we see between boroughs are statistically meaningful. The visualizations suggest that Brooklyn and Bronx have higher shooting counts than the other boroughs, but graphs alone cannot control for the changes in shootings over time. Since shootings rise and fall during different periods, especially around 2012–2018 and again in 2020, I use a simple Poisson regression model to account for the effect of year. This model allows me to check whether some boroughs still have significantly more shootings even after adjusting for the yearly trend. This helps verify that the borough differences are real and not just caused by variation across years.

```
# create yearly shooting counts per borough
shooting_counts = shootings_clean %>%
  mutate(YEAR = year(OCCUR_DATE)) %>%
  group_by(BORO, YEAR) %>%
  summarize(count = n(), .groups = "drop")
```

```r
#Set MANHATTAN as reference
shooting_counts$BORO <- relevel(
  factor(shooting_counts$BORO),
  ref = "MANHATTAN"
)

# Build the Poisson model
model = glm(count ~ BORO + YEAR, data = shooting_counts, family = poisson)

# Check model
summary(model)
```

```
##
## Call:
## glm(formula = count ~ BORO + YEAR, family = poisson, data = shooting_counts)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       53.842438   2.147099  25.077  < 2e-16 ***
## BOROBRONX          0.801413   0.019144  41.863  < 2e-16 ***
## BOROBROOKLYN       1.081260   0.018406  58.746  < 2e-16 ***
## BOROQUEENS         0.110307   0.021898   5.037 4.72e-07 ***
## BOROSTATEN ISLAND -1.579042   0.038470 -41.046  < 2e-16 ***
## YEAR              -0.024076   0.001066 -22.587  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16015.9  on 94  degrees of freedom
## Residual deviance:  1818.6  on 89  degrees of freedom
## AIC: 2516.1
##
## Number of Fisher Scoring iterations: 4
```

After controlling for year, the model shows clear differences between the boroughs. Brooklyn and the Bronx have significantly higher shooting counts compared to Manhattan, while Queens has only a slightly higher level. Staten Island remains much lower than all other boroughs. The coefficient for year is negative and significant, which means shootings have been decreasing over time across the city. These results confirm the patterns we observed in the visualizations and show that the borough differences are statistically meaningful, even after adjusting for yearly changes.

7. Conclusion

In this project, I analyzed the **NYPD Shooting Incident Data (Historic)** to explore the location patterns of shooting incidents in New York City from 2006 to 2024. By looking at the data from both the borough and precinct levels, I was able to identify where shootings happened most often and how these patterns changed over time. The descriptive results showed that Brooklyn and the Bronx consistently have the highest shooting activity. The top 5 precincts in each borough also revealed clear hotspots that contribute to the higher numbers in Brooklyn and the Bronx. To confirm the differences across boroughs, I used a simple Poisson regression model. After controlling for year, the model showed that Brooklyn and the Bronx still have significantly higher shooting counts than Manhattan. The year coefficient was negative, which supports the overall decreasing trend in recent years. This means the patterns we saw in the plots are not just due to certain unusual years but reflect real differences between the boroughs.

8. Possible Bias

There are some limitations in this analysis. The data only includes reported shootings, so it may miss incidents that were never reported or were recorded incorrectly. Reporting practices may also have changed over the years. Looking at boroughs also ignores differences in population size, neighborhood conditions, and police activity, which could affect the number of shootings in each area. Those limit the prescriptive value of my findings.

I also recognize that I might have my own personal bias when looking at this data. Before doing this project, I already had impressions about which boroughs were "more dangerous," and that could influence how I expect the results to look. To reduce this bias, I followed the same cleaning steps for all boroughs, used the data directly without making assumptions, and relied on the model output instead of my own expectations.